Formulário de Estatística Descritiva

Abrantes Araújo Silva Filho

2021-05-07 (v. 2)

Sumário

1 Descrição, exploração e comparação **2**
1.1 Medidas de tendência central 2
1.1.1 Para dados não agrupados 2
1.1.2 Para dados agrupados em classes sem intervalos 3
1.1.3 Para dados agrupados em classes com intervalos 3
1.2 Medidas de posição e separatrizes (quantis) 5
1.2.1 Observações importantes 5
1.2.2 Separatrizes (quantis) para dados não agrupados 6
1.2.3 Separatrizes (quantis) para dados agrupados em classes sem intervalos 7
1.2.4 Separatrizes (quantis) para dados agrupados em classes com intervalos 7
1.3 Medidas de dispersão 7
1.3.1 Para dados não agrupados 7
1.3.2 Para dados agrupados em classes sem intervalos 9
1.3.3 Para dados agrupados em classes com intervalos 10
1.3.4 Coeficiente de variação e escores z 12

1 Descrição, exploração e comparação

1.1 Medidas de tendência central

1.1.1 Para dados não agrupados

Média Aritmética:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Média Aritmética Ponderada:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

Média Geométrica:

$$\overline{x}_{\text{geom.}} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{3}$$

Média Harmônica:

$$\overline{x}_{\text{harm.}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{4}$$

Média Quadrática:

$$\overline{x}_{\text{quad.}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \tag{5}$$

Mediana:

$$\text{Mediana}(x) = \frac{1}{2}\left(x_{\left\lfloor \frac{(n+1)}{2} \right\rfloor} + x_{\left\lceil \frac{(n+1)}{2} \right\rceil}\right) \tag{6}$$

Moda:

$$\text{Moda}(x) = \text{valor(es) que mais se repete(m)} \tag{7}$$

Ponto Médio:

$$\text{Ponto médio}(x) = \frac{\max(x) + \min(x)}{2} \tag{8}$$

1.1.2 Para dados agrupados em classes sem intervalos

Média Aritmética Ponderada: basta calcular uma média ponderada onde os pesos w_i são as freqüências absolutas f_i de cada valor x_i (e k representa o número de classes/grupos)

$$\overline{x} = \frac{\displaystyle\sum_{i=1}^{k} x_i f_i}{\displaystyle\sum_{i=1}^{k} f_i} \tag{9}$$

Mediana: a mediana é calculada normalmente tendo-se o cuidado de utilizar as freqüências acumuladas para se determinar os valores corretos de $x_{\lfloor (n+1)/2 \rfloor}$ e de $x_{\lceil (n+1)/2 \rceil}$

$$\text{Mediana}(x) = \frac{1}{2}\left(x_{\left\lfloor \frac{(n+1)}{2} \right\rfloor} + x_{\left\lceil \frac{(n+1)}{2} \right\rceil} \right) \tag{10}$$

Moda: basta consultar a tabela de freqüências para determinar a(s) moda(s)

$$\text{Moda}(x) = \text{valor(es) que mais se repete(m)} \tag{11}$$

Ponto Médio:

$$\text{Ponto médio}(x) = \frac{\max(x) + \min(x)}{2} \tag{12}$$

1.1.3 Para dados agrupados em classes com intervalos

Obviamente os cálculos apresentados aqui são uma espécie de "gambiarra" pois o bom senso manda não realizar cálculos com dados agrupados em classes com intervalos pois os resultados serão sempre aproximados e incorretos (sempre que os dados são agrupados em intervalos estamos perdendo informações).

Na verdade eu nem iria incluir tais fórmulas aqui mas, como podem cair nas provas e concursos, mesmo contra minha vontade, seguem abaixo. Mas saiba que ninguém usa esse tipo de coisa, ninguém calcula nada sério com dados agrupados em intervalos de classe.

Média Aritmética Ponderada: basta calcular uma média ponderada onde os valores x_i serão representados pelo ponto médio do intervalo, PM_i, e os pesos w_i são as freqüências absolutas f_i de cada intervalo (k representa o número de classes/intervalos)

$$\overline{x} = \frac{\sum_{i=1}^{k} \text{PM}_i f_i}{\sum_{i=1}^{k} f_i} \tag{13}$$

Mediana: aqui a coisa fica um pouco mais complicada pois, em primeiro lugar, temos que encontrar a classe na qual a mediana se encontra e, depois, utilizar uma fórmula específica que leva em conta outras coisas. A fórmula está abaixo e as explicações a seguir:

$$\text{Mediana}(x) = \text{LI}_m + \frac{h_m}{f_m}\left(\frac{n}{2} - \text{FA}_{m-1}\right) \tag{14}$$

onde temos que:

- m é a classe na qual a mediana se encontra. A posição da mediana é obtida através do cálculo $\frac{1}{2}\left(\left\lfloor\frac{n+1}{2}\right\rfloor + \left\lceil\frac{n+1}{2}\right\rceil\right)$; com a posição calculada, basta consultar a tabela de freqüências acumuladas para identificar corretamente a classe m na qual a mediana se encontra.
- LI_m é o valor do limite inferior do intervalo da classe que contém a mediana.
- h_m é a amplitude do intervalo da classe que contém a mediana.
- f_m é a freqüência absoluta de observações no intervalo de classe que contém a mediana.
- FA_{m-1} é a freqüencia absoluta acumulada da classe imediatamente anterior ($m-1$) à classe da mediana (m).

Moda: aqui a coisa também fica um pouco mais complicada pois existem pelo menos 3 maneiras de se calcular a moda em dados agrupados com intervalos de classe, a moda bruta, a moda de King e a moda de Czuber. A moda bruta (equação 15) e a moda de Czuber (equação 16) são mostradas abaixo, com explicações a seguir:

$$\text{Moda bruta}(x) = \frac{\text{LI}_m + \text{LS}_m}{2} \tag{15}$$

$$\text{Moda de Czuber}(x) = \text{LI}_m + \left(\frac{\text{DA}}{\text{DA} + \text{DP}}\right) h \tag{16}$$

onde temos que:

- m: classe que contém a moda (basta ver a classe que a maior freqüência: essa é a classe que contém a moda).
- LI_m é o valor do limite inferior do intervalo da classe que contém a moda.
- LS_m é o valor do limite superior do intervalo da classe que contém a moda.
- DA é a diferença de freqüência entre a classe modal e a classe imediatamente anterior.
- DP é a diferença de freqüência entre a classe modal e a classe imediatamente posterior.
- h é a amplitude do intervalo de classe da classe que contém a moda.

1.2 Medidas de posição e separatrizes (quantis)

1.2.1 Observações importantes

O cálculo dos quantis (quartis, decis e percentis) é extremamente confuso entre os diversos autores e livros. Basta consultar alguns livros diferentes e você verá que existem quase tantos modos de se calcular os quantis quantos autores que os descreveram. E a coisa ainda fica pior quando você utiliza algum software para os cálculos porque:

1. Você pode não saber que método o software utiliza para o cálculo e, mesmo seguindo um exmplo de algum livro, o resultado que o software apresenta é um pouco diferente;
2. O software pode oferecer várias maneiras diferentes de calcular os quantis (existem mais de 10 maneiras em uso hoje em dia, das quais 8 são bem comuns) e você pode não saber como ajustar o software para reproduzir os resultados do livro que você está seguindo; e
3. Os softwares mais sofisticados (R, Octave, Mathematica, Matlab, Stata etc.), além de oferecerem vários modos diferentes de cálculo dos quantis, utilizam por padrão cálculos avançados, interpolações polinomiais e métodos que nós, meros mortais, dificilmente entendemos (figura 1). O melhor que você pode fazer é tentar aprender a utilizar o software do melhor modo possível, tentando mimetizar o que você está estudando nos livros, ou "abandonar" os cálculos mais simples (mas não tão

exatos) apresentados nos livros e tomar como "verdade" os resultados dos softwares (e mesmo assim, esse resultado pode variar de software para software, dependendo do método de cálculo utilizado internamente).

Figura 1: Alguns métodos de cálculos de quantis no Mathematica

- Common choices of parameters include:

{{0, 0}, {1, 0}}	inverse empirical CDF (default)
{{0, 0}, {0, 1}}	linear interpolation (California method)
{{1/2, 0}, {0, 0}}	element numbered closest to *qn*
{{1/2, 0}, {0, 1}}	linear interpolation (hydrologist method)
{{0, 1}, {0, 1}}	mean-based estimate (Weibull method)
{{1, -1}, {0, 1}}	mode-based estimate
{{1/3, 1/3}, {0, 1}}	median-based estimate
{{3/8, 1/4}, {0, 1}}	normal distribution estimate

FONTE: documentação do Mathematica
<https://reference.wolfram.com/language/ref/Quantile.html>

1.2.2 Separatrizes (quantis) para dados não agrupados

Percentil: aprendendo-se a calcular os percentis, pode-se calcular qualquer outro quantil (decis e quartis) pela equivalência entre eles. O cálculo apresentado aqui é bem simples e pode não coincidir com o obtido em softwares estatísticos (ver seção 1.2.1).

$$L_k(x) = \left(\frac{k}{100}\right) n; \quad \text{se} \begin{cases} L_k \in \mathbb{Z}, & P_k(x) = \dfrac{x_{L_k} + x_{L_k+1}}{2} \\ L_k \in \mathbb{R}, & P_k(x) = x_{\lceil L_k \rceil} \end{cases} \tag{17}$$

onde temos que:

- k é o percentil que queremos encontrar, em geral $1 \leq k \leq 99$.
- n é a quantidade de observações.
- L_k é o localizador do percentil k, ou seja, indicará a posição do elemento que será o k-ésimo percentil.
- P_k é o percentil k que queremos encontrar. Note que o cálculo dependerá se o localizador L_k é um número inteiro ou fracionário. Se for inteiro, o P_k será a média aritmética entre o valor na posição L_k e na imediatamente seguinte, $L_k + 1$. Se for fracionário, P_k será o valor de $x_{\lceil L_k \rceil}$, ou seja, o valor de x na posição teto de L_k.

1.2.3 Separatrizes (quantis) para dados agrupados em classes sem intervalos

Percentil: utiliza-se a equação 17 apresentada na seção 1.2.2, com o cuidado adicional de identificar os valores na freqüência acumulada da distribuição de freqüências.

1.2.4 Separatrizes (quantis) para dados agrupados em classes com intervalos

Percentil: aqui a coisa já complica mais um pouco mais complicada pois, em primeiro lugar, temos que encontrar a classe na qual a percentil desejado se encontra e, depois, utilizar uma fórmula específica que leva em conta outras coisas. A fórmula está abaixo e as explicações a seguir:

$$P_k(x) = \text{LI}_m + \frac{h_m}{f_m}\left[\left(\frac{k}{100}\right)n - \text{FA}_{m-1}\right] \tag{18}$$

onde temos que:

- k representa o percentil desejado.
- m é a classe no qual o percentil desejado se encontra. A posição do percentil é obtida através do cálculo $\left(\frac{k}{100}\right)n$; com a posição calculada, basta consultar a tabela de freqüências acumuladas para identificar corretamente a classe m no qual o percentil k se encontra.
- LI_m é o valor do limite inferior do intervalo da classe que contém o percentil.
- h_m é a amplitude do intervalo da classe que contém o percentil.
- f_m é a freqüência absoluta de observações no intervalo de classe que contém o percentil.
- FA_{m-1} é a freqüencia absoluta acumulada da classe imediatamente anterior $(m-1)$ à classe do percentil (m).

1.3 Medidas de dispersão

1.3.1 Para dados não agrupados

Amplitude:

$$\text{Amplitude}(x) = \max(x) - \min(x) \tag{19}$$

Desvio médio: a fórmula depende se estamos tratando do desvio médio populacional (equação 20) ou do desvio médio amostral (equação 21)

$$\text{Desvio médio}(X) = \frac{\sum_{i=1}^{N} |(X_i - \mu)|}{N} \tag{20}$$

$$\text{Desvio médio}(x) = \frac{\sum_{i=1}^{n} |(x_i - \overline{x})|}{n-1} \tag{21}$$

Variância: a fórmula depende se estamos tratando da variância populacional (σ^2, equação 22) ou da variância amostral (s^2, equação 23). Também existem fórmulas simplificadas para o cálculo (equações 24 e 25)

$$\sigma^2(X) = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \tag{22}$$

$$s^2(x) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \tag{23}$$

Variância (fórmulas simplificadas): existem fórmulas simplificadas que facilitam o cálculo da variância (equações 24 e 25)

$$\sigma^2(X) = \frac{N\left(\sum_{i=1}^{N} X_i^2\right) - \left(\sum_{i=1}^{N} X_i\right)^2}{N^2} \tag{24}$$

$$s^2(x) = \frac{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)} \tag{25}$$

Desvio-padrão: a fórmula depende se estamos tratando do desvio-padrão populacional (σ, equação 26) ou do desvio-padrão amostral (s, equação 27)

$$\sigma(X) = \sqrt{\sigma^2(X)} = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \tag{26}$$

$$s(x) = \sqrt{s^2(x)} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \tag{27}$$

Desvio-padrão (fórmulas simplificadas): existem fórmulas simplificadas que facilitam o cálculo do desvio-padrão (equações 28 e 29)

$$\sigma(X) = \sqrt{\sigma^2(X)} = \sqrt{\frac{N\left(\sum_{i=1}^{N} X_i^2\right) - \left(\sum_{i=1}^{N} X_i\right)^2}{N^2}} \tag{28}$$

$$s(x) = \sqrt{s^2(X)} = \sqrt{\frac{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}} \tag{29}$$

1.3.2 Para dados agrupados em classes sem intervalos

Amplitude: o cálculo é o mesmo, basta observar os valores na tabela de freqüências

$$\text{Amplitude}(x) = \max(x) - \min(x) \tag{30}$$

Variância: o cálculo é basicamente o mesmo, onde f_i representa a freqüência absoluta de cada X_i (ou x_i) e k representa o número de classes/grupos. Há fórmulas para a variância populacional e amostral.

$$\sigma^2(X) = \frac{\sum_{i=1}^{k} f_i(X_i - \mu)^2}{N} \tag{31}$$

$$s^2(x) = \frac{\sum_{i=1}^{k} f_i(x_i - \overline{x})^2}{n-1} \tag{32}$$

Variância (fórmulas simplificadas): existem fórmulas simplificadas que facilitam o cálculo da variância para dados agrupados sem intervalos de classe

$$\sigma^2(X) = \frac{N\left(\sum_{i=1}^{k} f_i X_i^2\right) - \left(\sum_{i=1}^{k} f_i X_i\right)^2}{N^2} \tag{33}$$

$$s^2(x) = \frac{n\left(\sum_{i=1}^{k} f_i x_i^2\right) - \left(\sum_{i=1}^{k} f_i x_i\right)^2}{n(n-1)} \tag{34}$$

Desvio-padrão: o cálculo é basicamente o mesmo, onde f_i representa a freqüência absoluta de cada X_i (ou x_i) e k representa o número de classes/grupos. Há fórmulas para o desvio-padrão populacional e amostral.

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{k} f_i (X_i - \mu)^2}{N}} \tag{35}$$

$$s(x) = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \overline{x})^2}{n-1}} \tag{36}$$

Desvio-padrão (fórmulas simplificadas): existem fórmulas simplificadas que facilitam o cálculo do desvio-padrão para dados agrupados sem intervalos de classe

$$\sigma(X) = \sqrt{\frac{N\left(\sum_{i=1}^{k} f_i X_i^2\right) - \left(\sum_{i=1}^{k} f_i X_i\right)^2}{N^2}} \tag{37}$$

$$s(x) = \sqrt{\frac{n\left(\sum_{i=1}^{k} f_i x_i^2\right) - \left(\sum_{i=1}^{k} f_i x_i\right)^2}{n(n-1)}} \tag{38}$$

1.3.3 Para dados agrupados em classes com intervalos

Amplitude: o cálculo leva em conta os limites do primeiro e do último intervalo

$$\text{Amplitude}(x) = \text{LS}_u - \text{LI}_p \tag{39}$$

onde temos que:

- LS_u é o valor do limite superior do último intervalo.
- LI_p é o valor do limite inferior do primeiro intervalo.

Variância: o cálculo leva em conta o ponto médio PM_i de cada classe. Novamente f_i representa a freqüência absoluta de cada classe e k representa o número de classes/grupos. Há fórmulas para a variância populacional e amostral.

$$\sigma^2(X) = \frac{\sum_{i=1}^{k} f_i(\text{PM}_i - \mu)^2}{N} \tag{40}$$

$$s^2(x) = \frac{\sum_{i=1}^{k} f_i(\text{PM}_i - \overline{x})^2}{n-1} \tag{41}$$

Variância (fórmulas simplificadas): existem fórmulas simplificadas que facilitam o cálculo da variância para dados agrupados com intervalos de classe

$$\sigma^2(X) = \frac{N\left(\sum_{i=1}^{k} f_i(\text{PM}_i)^2\right) - \left(\sum_{i=1}^{k} f_i\text{PM}_i\right)^2}{N^2} \tag{42}$$

$$s^2(x) = \frac{n\left(\sum_{i=1}^{k} f_i(\text{PM}_i)^2\right) - \left(\sum_{i=1}^{k} f_i\text{PM}_i\right)^2}{n(n-1)} \tag{43}$$

Desvio-padrão: o cálculo é basicamente o mesmo mas leva em conta o ponto médio PM_i de cada classe. Novamente f_i representa a freqüência absoluta de cada classe e k representa o número de classes/grupos. Há fórmulas para o desvio-padrão populacional e amostral.

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{k} f_i(\text{PM}_i - \mu)^2}{N}} \tag{44}$$

$$s(x) = \sqrt{\frac{\sum_{i=1}^{k} f_i(\text{PM}_i - \overline{x})^2}{n-1}} \tag{45}$$

Desvio-padrão (fórmulas simplificadas): existem fórmulas simplificadas que facilitam o cálculo do desvio-padrão para dados agrupados com intervalos de classe

$$\sigma(X) = \sqrt{\frac{N\left(\sum_{i=1}^{k} f_i(\text{PM}_i)^2\right) - \left(\sum_{i=1}^{k} f_i\text{PM}_i\right)^2}{N^2}} \tag{46}$$

$$s(x) = \sqrt{\frac{n\left(\sum_{i=1}^{k} f_i(\text{PM}_i)^2\right) - \left(\sum_{i=1}^{k} f_i\text{PM}_i\right)^2}{n(n-1)}} \tag{47}$$

1.3.4 Coeficiente de variação e escores z

Coeficiente de variação:

$$\text{CV}(X) = \frac{\sigma}{\mu} \times 100 \tag{48}$$

$$\text{CV}(x) = \frac{s}{\overline{x}} \times 100 \tag{49}$$

Escores z

$$z(X) = \frac{X_i - \mu}{\sigma} \tag{50}$$

$$z(x) = \frac{x_i - \overline{x}}{s} \tag{51}$$